STATISTICS AS A MATHEMATICAL DISCIPLINE

BY

D. V. LINDLEY

TECHNICAL REPORT NO. 33
APRIL 25, 1979

PREPARED UNDER GRANT
DAAG29-77-G-0031
FOR THE U.S. ARMY RESEARCH OFFICE

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

STATISTICS AS A MATHEMATICAL DISCIPLINE

By

D. V. Lindley

TECHNICAL REPORT NO. 33

April 25, 1979

Prepared under Grant DAAG29-77-G-0031

For the U.S. Army Research Office

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Partially supported under Office of Naval Research Contract N00014-76-C-0475 (NR-042-267) and issued as Technical Report No. 272.

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

STATISTICS AS A MATHEMATICAL DISCIPLINE

D. V. Lindley

(This paper is a revised version of a lecture given to a general mathematical audience at the First Australasian Mathematical Convention held in Christchurch, 15-19 May 1978. The revision consists of increasing the interpretative remarks at the cost of more mathematical ones; the judgment being that these will be of more interest to the readers of this journal. The example which concluded the lecture has been replaced by a different one, though emphasizing the same point, because the problem with which it deals has been discussed recently in Australasia, in particular by Davies (1978) at the conference. The example originally given in the lecture has appeared in Lindley and Phillips (1976). I am grateful for the award of an Erskine Fellowship by the University of Canterbury which enabled me to attend the convention and profit from a stay on the campus.)

STATISTICS AS A MATHEMATICAL DISCIPLINE

D. V. Lindley

1. The Axioms of Statistics.

Mathematicians who study statistics, perhaps through having to give a course on it, are nearly always surprised by what they find. (Actually "surprise" is perhaps the politest term used: the adjective is often derogatory.) This is reasonable, because they do not find a discipline akin to what they are used to in other branches of mathematics. They do not find a system that starts from fundamental notions incorporated in axioms, proceeding through definitions to theorems expressing important results in the field. Instead they find a collection of loosely related techniques, such as confidence intervals, significance tests; and a resulting confusion about what to do in any particular case. The purpose of this paper is to show that statistics can be regarded as a formal, mathematical system, like Euclidean geometry, so that mathematicians need no longer be "surprised".

Of course, I am not advocating this formalism just to keep mathematicians happy. That is a worthy objective, but not as important as the consideration that the mathematical method (if I can call it that, without being too precise about what is meant) works extremely well, and it is reasonable to expect advantages from applying it to statistical problems. With that method we know exactly what is true and what is false. In any problem we know what to do - in Newtonian mechanics write down the equations of motion - and thought can be concentrated on how to do it. Indeed,

the main advantage of this approach to statistics is that it gives practically useful results and is of just as much interest to the practitioner of statistics as to the mathematical statistician. We shall also look at an exciting theorem that is of great practical importance.

We begin, like Euclidean geometry, with the axioms. There is naturally freedom of choice here, but I shall take a system used by DeGroot (1970) in his excellent text. First it is necessary to say what is under discussion: the equivalent of the "points" and "lines" of geometry. In statistics we refer to <u>events</u> and we try to capture feelings of uncertainty that we have about events. For example, the event that a treatment will increase the yield. Statistical inference is a way of handling uncertainty. The events are denoted A, B, C, ... and they will be supposed to form a σ-field in a space S. This restriction merely means that the events can always be combined, to form a new event, in the usual ways. Our first axiom introduces the relation "not more likely than" between events, and is written $\underset{\sim}{\prec}$.

A1: For any two events A, B; either $A \underset{\sim}{\prec} B$, $B \underset{\sim}{\prec} A$ or both.

This says that any two events can be compared. In the usual way we obtain $A \sim B$, A and B are equally likely, and $A \prec B$, A is less likely than B. The second axiom shows what happens to the relation when events are combined in certain ways.

A2: If $A_1A_2 = B_1B_2 = \phi$ and $A_i \underset{\sim}{\prec} B_i$ $(i=1, 2)$ then $\cup A_i \underset{\sim}{\prec} \cup B_i$. Furthermore $A_1 \prec B_1$ and $A_2 \underset{\sim}{\prec} B_2$ imply $\cup A_i \prec \cup B_i$.

Here two events are each broken up into component events: if the components of one event are not more likely than the components of the other, then the same relation holds between the original events. It is easy to

deduce that these two axioms imply transitivity, namely $A \precsim B$ and $B \precsim C$ imply $A \precsim C$, and this may be taken as an axiom if preferred though it is weaker than the present one.

Other axioms follow, but these two are the key ones and it is worth making a few comments on them. First, the axioms are normative, not descriptive: that is, they refer to a norm or standard of behavior that you would like to achieve if you knew how; they do not describe your abilities at the moment. The task of statistics is to provide a satisfactory way of measuring uncertainty; the suggestion is that such measurements would obey the two requirements so far set out. Without statistics you may not be able to effect the required comparisons of events. Second, the axioms can be given an operational test in terms of bets: $A \prec B$ if a bet to win when B is true is preferred to a bet to win the same amount when A is true. This operational interpretation is important in order that the system can be used. Without it the system remains pure mathematics.

The third axiom relates $\precsim$ to the concept of inclusion, or implication, for events:

A3: $\phi \precsim A$ and $\phi \prec S$.

A consequence of this is that $A \subset B$ implies $A \precsim B$: if A implies B, then A is not more likely than B.

The fourth axiom is a matter of some contention. It is introduced in order to extend the concepts of the first three axioms from finite collections of events to enumerable collections. It is possible to dispense with it, but it is certainly simpler to include it.

A4: If $A_1 \supset A_2 \supset \ldots$ and $A_i \succsim B$ for all i, then $\cap A_i \succsim B$.

In words, if events get more and more restrictive but are never less likely than B, then the total restriction is not less likely than B. In technical language, this leads to σ-additivity.

Our object is to measure uncertainty; that is, to associate with each event a number describing that uncertainty. It had been conjectured that these four axioms would be enough to do this, but a counterexample showed that this was not so. They have therefore to be strengthened. There are various ways of doing this, but all of them essentially amount to introducing a standard of uncertainty to which all events can be referred. Indeed, in any measurement process, say of length or mass, measurement is always with respect to a standard; so it is by no means unusual to introduce the same concept in connection with uncertainty. It is simplest to invoke a strong standard that brings continuity along with it. This we do by supposing the σ-field contains the Borel sets of the unit interval $[0, 1]$ of the real line and presume:

A5: There exists a uniform probability distribution in $[0, 1]$. This is our standard. By a uniform distribution is meant an assignment of uncertainty such that if I_1 and I_2 are any two intervals in $[0, 1]$ of equal length then they are judged equally likely, $I_1 \sim I_2$.

It is now a straightforward matter to use a Dedekind-type of argument to show that for any A, there exists an interval I, only the length of which matters, such that $A \sim I$. With each A is associated a number, equal to the length of the corresponding I and called the probability of A. Furthermore it can be shown that these numbers obey the usual laws of probability discussed below, and therefore are entitled to the name probability. These probabilities are written $p(A|S)$ - read "the probability of A given S" - since they depend on S as well as the uncertain event under consideration. An example will demonstrate the truth of this latter point. Suppose S contains the $(n+1)$ integers

(0, 1, 2, ..., n) corresponding to the number of heads in n tosses of a coin: then your probability of A, exactly 2 heads, will be altered if S is restricted to include only the even integers.

Our final axiom is concerned with such a restriction of S to a subset C say, with $C \succ \phi$. We introduce a notion of A is not more likely than B, given C and write $(A|C) \underset{\sim}{\prec} (B|C)$, the original form obtaining when C=S. These ordering relations are connected with the original ordering by

A6: For any A,B and C, $C \succ \phi$, $(A|C) \underset{\sim}{\prec} (B|C)$ iff $AC \underset{\sim}{\prec} BC$. The motivation behind this assumption is best explained in terms of the operational bets referred to above. Suppose that you are contemplating a bet which will only take place if C occurs and then gives a reward only if A does, and are comparing it with a second bet under the same conditions except that the reward hinges on B. Then the rewards will only arise if A and C, or B and C, both occur. Consequently the relevant uncertainties concern AC and BC. The bets are "called-off" if C does not occur, so this is sometimes referred to as the axiom of called-off bets.

These six axioms complete the system and from them it is possible to prove the basic

Theorem: A1 - A6 imply that there exists a unique probability distribution $p(A|B)$ for all A, and all $B \succ \phi$, such that $(A|C) \underset{\sim}{\prec} (B|C)$ iff $p(A|C) \leq p(B|C)$.

By a probability distribution is meant an assignment of numbers, written $p(A|B)$, to all events A, and all events B with $p(B|S) > 0$ such that

P1: $p(A|A) = 1$ and $0 \leq p(A|B) \leq 1$. (Convexity)

P2: If $\{A_i: i=1, 2, \ldots\}$ are exclusive, $A_iA_j = \phi$ for all unequal i,j, then $p(\cup A_i|B) = \Sigma p(A_i|B)$. (Additivity)

P3: $p(AB|C) = p(A|C)\ p(B|AC)$. (Multiplicativity)

(The three properties are often described by the names given in brackets.)

2. The Likelihood Principle

In words, the theorem says that an appreciation of uncertainty within the framework of the axioms A1 - A6 is only possible through the notion of probability: uncertainty can only be described probabilistically. To use any other method that is not equivalent to a probability calculus will lead to violation of one of the axioms. Would you wish to do that? This is a striking result since, as we shall see below, statisticians do use other methods and get into difficulties as a result. Since the proof of the theorem is simple, the only objection to the approach can be in the axioms: are they satisfactory? The one that is most obviously open to attack is A4 with its possibly infinite set of events; but if it is omitted P1 - P3 remain except that in P2 the events can only be finite in number and the probability is finitely-, and not σ-, additive. This leads to difficulties: for example, the marginalization paradoxes of Dawid, Stone and Zidek (1973). Another reason for thinking that the system is satisfactory is that it is possible to approach the topic of uncertainty in other ways and still end up with P1 - P3 or variants thereof. We cite the work of DeFinetti (1974). More modest approaches that deal with special systems lead to the likelihood principle that we will discuss below: see, for example, Birnbaum (1962) or Basu (1975). Another aspect of this approach to uncertainty is through bets, as already mentioned.

It is possible to show that any measurement of uncertainty that can be used as a basis for bets and is not probabilistic results in a Dutch book: that is a combination of bets that will lose you money for sure, whatever happens. Dutch books can be made for most statistical practices.

The argument is like that used in other branches of mathematics. Furthermore it is operational. One can test people by means of bets to see if they react in the way the results require: it is good applied, as well as pure, mathematics. Furthermore we are now in a position to prove other theorems and can test these against practical experience. One such theorem is Bayes theorem which says that whenever the probabilities exist $p(B_i|AC) \propto p(A|B_iC)\, p(B_i|C)$ for $i=1, 2, \ldots$, the omitted constant of proportionality not depending on i. This theorem plays such an important role in modern statistics that methodology based on the axioms is often called Bayesian statistics. It is better to refer to the axioms as those of coherence because they typically deal with how judgments of uncertainty get together, or cohere; so that perhaps the appropriate description is coherent statistics. But Bayes theorem does have an astonishing consequence for statistics that we now explore.

In statistical problems S is a product space $\mathcal{X} \times \Theta$ of elements (x,θ). (For simplicity we will describe the situation where S is enumerable. The results generalize to more general classes of spaces.) The quantity x is referred to as the data and θ as the parameter. The probabilities, given S, are most easily described by a density $p(x,\theta)$, never negative and with $\Sigma_{x,\theta}\, p(x,\theta) = 1$. Then $p(A|S) = \Sigma_{(x,\theta)\in A}\, p(x,\theta)$ and generally $p(A|B)$ is $p(AB|S)/p(B|S)$ by P3. In particular $p(\theta) = \Sigma_x\, p(x,\theta)$ and $p(x|\theta) = p(x,\theta)/p(\theta)$

provided $p(\theta) \neq 0$, which will henceforth be supposed true. (Conventional statistics would admit $p(x|\theta)$ but not $p(\theta)$.) The point of writing S in this way is that the statistician observes the value of x, but not of θ, and wishes to express his uncertainty about θ in the light of the observation. Bayes theorem allows him to calculate this as

$$p(\theta|x) \propto p(x|\theta)\ p(\theta)$$

the constant not involving θ. (In fact it is $p(x)^{-1}$.)

At this point we had better improve on our rather sloppy notation in order to emphasize the nature of the last result. Bayes theorem would better be written

$$p_\theta(\cdot|x) \propto p_x(x|\cdot)\ p_\theta(\cdot)$$

to emphasize the fact that x is known and that the functions are all functions of θ. The first and last of the three functions are clearly probability densities, never negative and adding to 1 ($\Sigma_\theta\ p_\theta(\theta|x) = \Sigma_\theta\ p_\theta(\theta) = 1$.) The other, $p_x(x|\cdot)$, is nonnegative but does not typically add to 1: it is not a probability (as a function of θ) and is termed the <u>likelihood</u> of θ, given x. In words the above result is described by saying that the probability of θ given x is proportional to the product of the likelihood of θ given x and the probability of θ prior to x. We have proved the following

<u>Theorem</u>. The uncertainty about θ, given x, depends on x only through the likelihood function $p_x(x|\cdot)$.

Expressed differently, if two data values, x_1 and x_2, have the same likelihood, then the uncertainty about θ, given x, is the same

as that given x_2. Or, in another form, the likelihood function is a sufficient statistic. It is often referred to as the <u>likelihood principle</u>.

This is a surprising result. To appreciate the reason for the astonishment, recall that it has been proved on the basis of some reasonable axioms about uncertainty, so should reasonably apply in practical situations of uncertainty. In fact, almost all methods used in statistics violate the principle. It is easy to see this since the only probability admitted in sampling theory statistics is $p_x(\cdot|\theta)$ for each θ and the procedures used involve integrals over x-values. For example, a significance test has $\int_R p_x(x|\theta)dx = \alpha$ over the rejection region R for values of θ constituting the null hypothesis; an unbiased estimate $t(x)$ employs $\int t(x)\, p_x(x|\theta)dx = \theta$ for all θ. In both cases, as in others, values of x besides that observed are used, so violating the likelihood principle. The method of maximum likelihood is an obvious exception, but that is unsatisfactory for a different reason that will appear below.

We now have a wonderful way of testing the ideas developed in this paper. According to Popper, a theory is valuable if several testable results can be deduced from it. And the theory fails as soon as a test fails. Our ideas are certainly valuable in this sense because all the results of the rich, probability calculus are available. And here we have a test: which ideas are better, those based on the likelihood principle, or those on conventional statistics? There is no room here to explore this question generally, so we confine ourselves to an example.

3. Inference for a Galton-Watson Process

Let us explore the likelihood principle in a situation that is of current research interest. Consider a Galton-Watson process in which at each generation, each individual alive then gives rise, independently of all other individuals, to a number r of offspring which constitute the next generation. Suppose that the probability density for r is the same for all individuals and is known up to an unknown parameter θ; write it $p_r(r|\theta)$. For simplicity suppose r is never zero so that extinction is not possible: the conclusion is unaffected if this restriction is removed. Since our object is to illustrate the differences between the likelihood approach and sampling theory ideas, and not to carry out detailed calculations, let us specialize to the case of the geometric distribution with $p_r(r|\theta) = (1-\theta)\theta^{r-1}$ for $r \geq 1$ and $0 < \theta < 1$.

Now suppose the process is observed for N generations starting with a single individual in the first generation. Let the numbers of offspring observed be r_{11} from that first generation individual, $r_{21}, r_{22}, \ldots, r_{2r_{11}}$ from the r_{11} at the second generation; and so on up to $r_{N1}, r_{N2}, \ldots, r_{NS}$ where S is the total at the $(N-1)$-generation. The probability for this data set $x = (r_{11}: r_{21}, r_{22}, \ldots, r_{2r_{11}}: r_{31}, \ldots: r_{N1}, \ldots, r_{NS})$ is clearly

$$(1-\theta)\theta^{r_{11}-1}\ (1-\theta)\theta^{r_{21}-1}\ (1-\theta)\theta^{r_{22}-1}\ \ldots\ (1-\theta)\theta^{r_{2r_{11}}-1}\ (1-\theta)\theta^{r_{31}-1}$$

$$\ldots\ (1-\theta)\theta^{r_{NS}-1}$$

or simply $(1-\theta)^{R_{N-1}}\,\theta^{R_N - R_{N-1}}$ where R_i is the total number of individuals up to and including the i^{th} generation. Essentially each

parent contributes a term $(1-\theta)/\theta$; each offspring a term θ. The total number of parents is R_{N-1} (the parenthood of those at the N^{th} generation is not observed) and of offspring R_N-1, since the original individual was not observed as an offspring. This is the likelihood function, and according to the likelihood principle no other aspect of the data is relevant to the uncertainty about θ and (R_N, R_{N-1}) is a sufficient statistic. (Notice it does not involve N, the number of generations.)

Another, more interesting, feature of the likelihood is that it is the same as that for a random sample of size R_{N-1} from the geometric distribution that yields a sum R_N-1. To see this, note that a random sample of size n with values $r_1, r_2, \ldots, r_n$ gives a likelihood $(1-\theta)^n \theta^{\Sigma(r_i-1)}$: writing $n = R_{N-1}$ and $\Sigma r_i = R_N-1$ gives the result. Consequently we have two data sets (x_1 and x_2 in an earlier notation) which have the same likelihood function and therefore for which the uncertainties about θ should be the same according to the principle: one data set observes for a fixed number N of generations, the other observes for a fixed sample size n. Now the latter situation is standard in the statistical literature and, for example, the maximum likelihood estimate of θ, $\Sigma(r_i-1)/\Sigma r_i$, is asymptotically normal with mean equal to the true value of θ and variance $\theta(1-\theta)^2/n$, the latter being obtained from the inverse of the expectation of minus the second derivative of the log-likelihood. According to the likelihood principle the same asymptotic inference as stated in the last sentence should be available for the original case of fixed generation number. Is this so?

The maximum likelihood example remains as before; in the changed notation it is $(R_N-R_{N-1}-1)/(R_N-1)$; but the variance is altered because

the expectation, which previously treated n as fixed and only Σr_i as random, now has <u>both</u> these quantities, involving R_N and R_{N-1}, as random. Detailed calculation shows that the expectation of minus the second derivative of the log-likelihood when inverted gives $(1-\theta)^N \theta^2/\{1-(1-\theta)^N\}$, or asymptotically $(1-\theta)^N \theta^2$, of quite different form from the earlier expression $\theta(1-\theta)^2/n$. Consequently the likelihood principle is in direct conflict with conventional statistical practice that uses a sampling variance to judge the imprecision of an estimate. The reason is not hard to see. The computation of the variance involves an integration of $p_x(x|\theta)$ over a suitable space which is different according to whether the generation number N or the number n of parents is fixed. The reader can judge for himself which seems sensible, remembering that if he favours the sampling-variance approach he is somewhere violating one of the axioms described above. For my part, the likelihood approach seems clearly correct: we have here R_{N-1} parents independently producing offspring, why should it matter that R_{N-1} is random or fixed? Another related difficulty with the sampling-theoretic approach is that in many cases it is hard to be sure what the relevant space for integration is. For example, suppose the N generations had been observed but that some parents had left the system during the period of observation so that there was no knowledge about their offspring. The likelihood remains $(1-\theta)^P \theta^{Q-P}$ where P and Q are respectively the number of parents and the number of offspring. In the original case of complete observations it is particularly curious to fix N since it is not even part of the sufficient statistic.

In this paper we have shown that it is possible to present statistics within the framework of a simple axiom-system capturing the concept of uncertainty; that this system leads to the result that uncertainty must be described probabilistically; that one probability result, Bayes theorem, leads to the likelihood principle, and that this principle is in direct conflict with statistical practice. We therefore have a piece of theory that cannot be ignored by the most applied of statisticians because it strongly affects practice.

REFERENCES

BASU, D. (1975). Statistical information and likelihood. Sankhyā A 37, 1-71.

BIRNBAUM, ALLAN (1962). On the foundations of statistical inference. J. Amer. Statist. Assn. 57, 269-306.

DAWID, A. P., STONE, M., and ZIDEK, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference. J. Roy. Statist. Soc. B 35, 189-233.

DeFINETTI, B. (1974). Theory of Probability. Two volumes. Wiley: New York.

DeGROOT, M. H. (1970). Optimal Statistical Decisions. McGraw-Hill: New York.

LINDLEY, D. V., and PHILLIPS, L. D. (1976). Inference for a Bernoulli process. The American Statistician 30, No. 3, 112-118.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: 33	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): STATISTICS AS A MATHEMATICAL DISCIPLINE		5. TYPE OF REPORT & PERIOD COVERED: TECHNICAL REPORT
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): D. V. LINDLEY		8. CONTRACT OR GRANT NUMBER(s): DAAG29-77-G-0031
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics, Stanford University, Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: P-14435-M
11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office, Post Office Box 12211, Research Triangle Park, NC 27709		12. REPORT DATE: APRIL 25, 1979
		13. NUMBER OF PAGES: 14
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report): UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents. This report partially supported under Office of Naval Research Contract N00014-76-C-0475 (NR-042-267) and issued as Technical Report No. 272.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Axioms for uncertainty, Normative, Likelihood principle, Coherence, Galton-Watson process

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

PLEASE SEE REVERSE SIDE

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

STATISTICS AS A MATHEMATICAL DISCIPLINE

It is argued that statistical inference, like other branches of mathematics, should have a structure of axioms and theorems. Such an axiomatic system is described and shown to lead to the likelihood principle. Almost all standard statistical techniques violate that principle. The paper concludes with an example using the Galton-Watson process which demonstrates the power of the principle.